Molecular Evolution of the *ocnus* and *janus* Genes in the *Drosophila melanogaster* Species Subgroup

John Parsch, Colin D. Meiklejohn,* Elisabeth Hauschteck-Jungen,† Peter Hunziker,‡ and Daniel L. Hartl**

*Department of Organismic and Evolutionary Biology, Harvard University; and †Zoologisches Institut and ‡Biochemisches Institut, Universität Zürich, Zürich, Switzerland

Genes involved in male fertility are potential targets for sexual selection, and their evolution may play a role in reproductive isolation and speciation. Here we describe a new *Drosophila melanogaster* gene, *ocnus* (*ocn*), that encodes a protein abundant in testes nuclear extracts. RT-PCR indicates that *ocn* transcription is limited to males and is specific to testes. *ocn* shares homology with another testis-specific gene, *janusB* (*janB*), and is located just distal to *janB* on chromosome 3. The two genes also share homology with the adjacent *janusA* (*janA*) gene, suggesting that multiple duplication events have occurred within this region of the genome. We cloned and sequenced these three genes from species of the *D. melanogaster* species subgroup. Phylogenetic analysis based on protein-encoding sequences predicts a duplication pattern of *janA* → *janA janB* → *janA janB ocn,* with the latter event occurring after the divergence of the *D. melanogaster* and *Drosophila obscura* species groups. We found significant heterogeneity in the rates of evolution among the three genes within the *D. melanogaster* species subgroup as measured by the ratio of nonsynonymous to synonymous substitutions, suggesting that diversification of gene function followed each duplication event and that each gene evolved under different selective constraints. All three genes showed faster rates of evolution than genes encoding proteins with metabolic function. These results are consistent with previous studies that have detected an increased rate of evolution in genes with reproductive function.

Introduction

Gene duplication is thought to play a fundamental role in the evolution of DNA sequences and in the creation of novel genetic material on which natural selection can act (Ohno 1970). Following a duplication event, the two paralogous genes may evolve under different selective constraints and thus may show different patterns of molecular evolution. For example, one gene may retain its original function and continue to evolve under the same selective constraints as before the duplication while the other gene accumulates mutations according to the spontaneous mutation rate of the organism and becomes a nonfunctional pseudogene. A hallmark of pseudogenes that indicates the lack of selective constraint on their sequence is that the three codon positions show nearly identical rates of nucleotide substitution. In species with a high rate of DNA deletion, pseudogenes may be rapidly lost or rendered unidentifiable. Such a process is hypothesized to explain the lack of bona fide pseudogenes in Drosophila (Petrov, Lozovskaya, and Hartl 1996).

A second possibility is that both copies of the duplicated gene retain the original function. Under this scenario, the two genes remain under the same selective constraints and should show similar patterns of molecular evolution. The maintenance of two redundant genes is not expected to be stable over evolutionary time unless accompanied by some breaking of symmetry, such as the partitioning of gene function (Force et al. 1999; Krakauer and Nowak 1999). On a shorter timescale, two redundant genes may persist if there is a selective advantage to having multiple copies of the same gene, as is proposed for the *Drosophila melanogaster* metallothionein gene, *Mtn.* Metallothioneins play an important role in the detoxification and intracellular regulation of heavy metals. Polymorphism for tandem duplication of *Mtn* has been found in *D. melanogaster,* and flies with the duplication show increased levels of *Mtn* expression (Lange, Langley, and Stephan 1990; Theodore, Ho, and Maroni 1991). The increased expression may be favored in environments exposed to heavy metal pollution over recent human history (Lange, Langley, and Stephan 1990).

A third possibility is that one gene copy may retain the original function while the other evolves a new function through changes to its amino acid sequence and/or expression pattern. In such a case, the gene adopting a new function is expected to experience selective pressures different from those experienced by the original gene. This is exemplified by the *Adh* and *Adhr* genes of Drosophila. The *Adh* gene product performs a well-known enzymatic function as an alcohol dehydrogenase. The function of the *Adhr* product is unknown, although considerable evidence suggests that it is not an alcohol dehydrogenase (reviewed by Ashburner 1998). Conservation of the *Adhr* coding sequence between *D. melanogaster* and *Drosophila pseudoobscura,* however, implies strong selective constraints and a functional role (Schaeffer and Aquadro 1987). Consistent with the above expectation, *Adh* and *Adhr* show different patterns of interspecific divergence and also differ in levels of codon bias (Schaeffer and Aquadro 1987; Albalat, Marfany, and Gonzalez-Duarte 1994). A more striking example of a gene duplication leading to a gene of novel function is the *D. melanogaster Sdic* gene. *Sdic* arose

Key words: *janus, ocnus,* gene duplication, sperm protein, *Drosophila melanogaster* species subgroup.

Address for correspondence and reprints: John Parsch, Department of Organismic and Evolutionary Biology, Harvard University Biological Laboratories, 16 Divinity Avenue, Cambridge, Massachusetts 02138-2020. E-mail: jparsch@oeb.harvard.edu.

Mol. Biol. Evol. 18(5):801–811. 2001

© 2001 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

by a duplication of the cytoplasmic dynien intermediate chain gene, *Cdic,* followed by fusion to the 5′ end of the *AnnX* gene and other rearrangements to produce a functional open reading frame (Nurminsky et al. 1998). *Sdic* has evolved a sperm-specific pattern of expression and appears to have become fixed rapidly in *D. melanogaster* due to the action of positive selection in a ''selective sweep'' (Nurminsky et al. 1998).

In this paper, we describe the molecular evolution of a newly identified sperm-specific gene, *ocnus* (*ocn*), and two related genes, *janusA* (*janA*) and *janusB* (*janB*). The three genes are arranged in tandem over a genomic region of less than 2.5 kb and appear to be the result of two separate duplication events. The *janB* gene also shows sperm-specific expression, while *janA* has two major alternatively spliced forms, one specific to the male germ line and the other showing a more general pattern of expression in both males and females (Yanicostas, Vincent, and Lepesant 1989). Patterns of molecular evolution in these genes may reflect their role in male fertility. It has been suggested that genes influencing male reproductive traits evolve rapidly, and this ''faster male'' hypothesis may provide a partial explanation for Haldane's rule in species where males are the heterogametic sex (Wu and Davis 1993; Presgraves and Orr 1998). A previous study comparing genes between *D. melanogaster* and either *Drosophila simulans* or *D. pseudoobscura* that included *janA* and *janB* found a higher ratio of nonsynonymous to synonymous substitution rates in genes with sex-related functions than in genes with metabolic functions (Civetta and Singh 1998). Here, we used protein-encoding sequences of the *ocn, janA,* and *janB* genes in the *D. melanogaster* species subgroup to examine the pattern of duplication within this genomic region and to compare selective constraints among the three genes.

Materials and Methods

Protein Purification and Sequencing

Protein isolation from *D. melanogaster* males was performed principally according to the method of Gusse et al. (1986), with a number of alterations to adapt it to Drosophila. Males of a wild-caught strain from Zürich, Switzerland, were kept isolated from females for at least 30 days at 18°C. The testes of 285 males were dissected in a solution containing 50 mM Tris-HCl (pH 7.0) and 2 mM EDTA. Tissue was collected in a solution of 50 mM Tris-HCl (pH 7.5), 2 mM EDTA, 1% diisopropylfluorophosphate (DPF), 0.5 mM p-chloromercuriphenylsulfonic acid (PCMPS), and 0.1% Nonidet. The testes were stored in this solution at −70°C. After warming, the tissue was sonified and sedimented by ultracentrifugation through a solution of 1.8 M sucrose, 20 mM Tris-HCl (pH 7.5), and 150 mM KCl. The pellet was suspended in 50 mM Tris-HCl (pH 7.5), 0.35 M NaCl, 0.8 M sucrose, 2 mM EDTA, 1% DPF, and 0.5 mM PCMPS for 1 h and centrifuged. The pellet was then washed with ethanol. For the first step of protein extraction, the pellet was incubated in 0.25 M HCl overnight at 4°C. After centrifugation, the soluble proteins in the supernatant were precipitated with 100% TCA to a final concentration of 20% and centrifuged, and the pellet was washed with acid acetone (100 ml acetone plus 0.05 ml concentrated HCl) and acetone and then dried. The pellet remaining after the first protein extraction was reduced with 0.2 M dithiothreitol in 8 M urea, 0.5 M NaCl, and 0.1 M Tris-HCl (pH 8.5) under N_2 at 37°C for 1 h and alkylated with 62.5 mM iodacetamide for another hour under N_2 at 37°C. For the second extraction step, 6 M HCl was added to the supernatant to a final concentration of 0.25 M and incubated overnight at 4°C. After centrifugation, the soluble proteins were precipitated with 100% TCA to a final concentration of 20% and centrifuged, and the pellet was washed with acid acetone and acetone and then dried. For a control sample, total protein was prepared from 12 one-day-old males using the first extraction step described above.

Proteins were analyzed by electrophoresis on an acetic acid/urea/polyacrylamide gel with 17% polyacrylamide and 6.25 M urea. The gel was stained with Coomassie brilliant blue. For sequence analysis, bands were transferred to a polyvinylidene difluoride membrane by electroblotting with 0.01 M acetic acid. The membrane was stained with amido black, and the appropriate band was cut out. Sequence analysis was carried out on a Model 477 sequencer (Applied Biosystems Inc., Foster City, Calif.) according to the manufacturer's recommendations.

Fly Stocks and Genomic DNA Preparation

A laboratory stock of the Canton S strain of *D. melanogaster* was used for all subsequent experiments. For *D. simulans, Drosophila yakuba,* and *Drosophila teissieri,* we used isofemale stocks derived from wild-caught flies. *Drosophila sechellia, Drosophila mauritiana, Drosophila erecta,* and *Drosophila orena* stocks were obtained from the Drosophila Species Stock Center (Bowling Green, Ohio). Genomic DNA was prepared from individual male flies by homogenization, followed by a 2-h incubation at 37°C in buffer (0.2 M sucrose, 0.1 M NaCl, 50 mM Tris-HCl [pH 8.0], 10 mM EDTA [pH 8.0], 0.5% Triton X-100) containing 1% sarkosyl and 50 μg/ml Proteinase K. After incubation, the homogenate was extracted twice with phenol : chloroform and once with chloroform. DNA was precipitated in 100% ethanol, washed in 70% ethanol, vacuum-dried, and resuspended in 1 × TE (pH 8.0).

PCR Cloning and DNA Sequencing

Unless otherwise noted, PCR and RT-PCR reagents were supplied by Life Technologies (Gaithersburg, Md.). Approximately 100 ng of genomic DNA template was amplified for 25 cycles (94°C for 1 min, 55°C for 1 min, 72°C for 3 min) in a 50-μl reaction containing 1 × PCR buffer, 2.5 mM magnesium chloride, 125 μM each dNTP, 100 ng each primer, and 1 U *Taq* DNA polymerase. The primers used for amplification were ja1 (5′-GTATCTGGTCACATTGCTGGAC-3′), ja2 (5′-GCAAAGCTACAGACTAACTGC-3′), jb1 (5′-GCAGTTAGTCTGTAGCTTTGC-3′), jb2

(5′-CC GAA AAG AAAC TGG TA TGAACGG-3′), oc1 (5′-CCGTTCATACCAGTTTCTTTTCGG-3′), and oc2 (5′-GGCAAGATGATGTTGTAATGCTGG-3′). For *D. melanogaster, D. simulans,* and *D. teissieri,* the *janA, janB,* and *ocn* genes were amplified separately using the primer pairs ja1-ja2, jb1-jb2, and oc1-oc2, respectively. For *D. yakuba, janA* was amplified with the ja1-ja2 primers, and the *janB-ocn* region was amplified as a 1.6-kb fragment using the jb1-oc2 primers. For *D. mauritiana, D. erecta,* and *D. orena,* the entire *janA-janB-ocn* 2.4-kb genomic region was amplified using the primers ja1-oc2. For *D. sechellia,* amplification with the ja1-ja2 primers was unsuccessful; however, we were able to amplify the 1.6-kb genomic region containing the *janB* and *ocn* genes using the primers jb1-oc2. PCR products were cloned with the TOPO TA Cloning kit (Invitrogen, Carlsbad, Calif.). Plasmid DNA was purified following the alkaline lysis protocol of Sambrook, Fritsch, and Maniatis (1989, pp. 1.25–1.28), with an additional 100% chloroform extraction performed before ethanol precipitation. Approximately 200 ng of plasmid DNA template was used per sequencing reaction using the Dye Terminator cycle sequencing kit (Applied Biosystems Inc.). Gene-specific PCR primers (listed above) and universal M13 forward and reverse primers were used as sequencing primers. An additional sequencing primer was designed specific to the *D. yakuba ocn* sequence (5′-CTGGTTAGGCCGTGCATGTG-3′). Sequencing gels were run on an ABI 373 automated sequencer. DNA was sequenced on both strands, and at least two independent clones were sequenced for each gene in each species. Additional independent clones were sequenced when necessary to resolve ambiguities.

RNA Preparation and RT-PCR

Total RNA was prepared from whole adult flies, adult body segments, or hand-dissected tissues using TRIzol reagent and following the manufacturer's protocol. In all cases, adult flies were collected 6–8 days after eclosion. For whole flies, separate RNA preparations were made using 10 flies of each sex. An additional 10 adult males were sectioned into head, thorax, and abdomen segments, with each segment being used for a separate RNA preparation. The testes and midguts from 40–50 adult males were isolated by hand dissection. These tissues were stored in RNAlater solution (Ambion, Austin, Tex.) prior to RNA extraction. First-strand cDNA was synthesized from approximately 500 ng of total RNA in a 20-μl reaction containing: 1 × first-strand buffer, 3 μg random primers, 0.1 M DTT, 1 U RNasin (Promega, Madison, Wis.), and 200 U Superscript II reverse transcriptase. The reaction was incubated for 1 h at 37°C and then heated to 65°C for 10 min. Five microliters of each cDNA reaction was used for PCR in a 50-μl reaction under the same conditions given above for genomic DNA amplification. As a control to ensure that sufficient cDNA was present, two PCR reactions were performed on each cDNA preparation. The two reactions were identical except for the primers. The first reaction contained primers specific to the *ocn* gene, while the second reaction contained primers specific to the actin gene, *Act5C* (GenBank accession number K00667). The *ocn* primers (oc1 and oc2) are expected to amplify a cDNA product of 489 bp; the *Act5C* primers (5′-GTGACGAAGAAGTTGCTGCTC-3′ and 5′-ATCTGCTGGAAGGTGGACGAC-3′) are expected to amplify a product of 1,063 bp. Following amplification, PCR products were separated on a 1% agarose gel and visualized under UV light by ethidium bromide staining.

Sequence Analysis

Protein-encoding sequences were aligned by first aligning amino acid sequences with the CLUSTAL X program (Thompson et al. 1997), then back-translating the aligned amino acids into codons. Slight adjustments to improve the alignment of the codons were made by eye. Gene trees were constructed by maximum parsimony using PAUP* (Swofford 2000) with 1,000 bootstrap replicates. Maximum-likelihood analyses of substitution rates were performed using the PAML software package (Yang 2000). Transition/transversion rates and nucleotide frequencies at the three codon positions were estimated separately for each gene except in the combined analysis, where the average values from all three genes were used for both the combined data and the individual genes.

Results

Identification of the *ocn* Gene

Electrophoresis of nuclear proteins extracted from the testes of mature males revealed a prominent band with a migration pattern similar to that of histones H2A and H2B (fig. 1). We obtained three partial amino acid sequences from this protein: DRVNALLINVPXV(T/Q)LLT, QTDLLLSWTR, and FQHGLADLFPK. The longest sequence was used in a BLAST search against an all-frame translation of expressed sequence tags (ESTs) in the Berkeley Drosophila Genome Project (BDGP) database (Rubin et al. 2000). A match of 14 consecutive residues was found to an EST (accession number GH02250). The second partial amino acid sequence matched this EST at 6 of the 8 residues (including five consecutive residues). The third partial sequence matched at 6 of 10 residues when a single gap was allowed. The longest consecutive match to this sequence was, however, only three residues. No other matches to any of our partial sequences were found in the BDGP EST database or in the database of predicted proteins generated from the complete *D. melanogaster* genome sequence (Adams et al. 2000). Although none of our partial sequences produced a perfect match, the combined results presented above strongly suggest that the protein we isolated from testes nuclear extracts is encoded by the GH02250 EST sequence. This conclusion is further supported by the observation that the predicted protein from this EST shares the greatest homology (43%) with the *D. melanogaster janB* protein, which is expressed specifically in testes. We designate the new gene *ocnus,* after the grandson of the Roman god Janus.

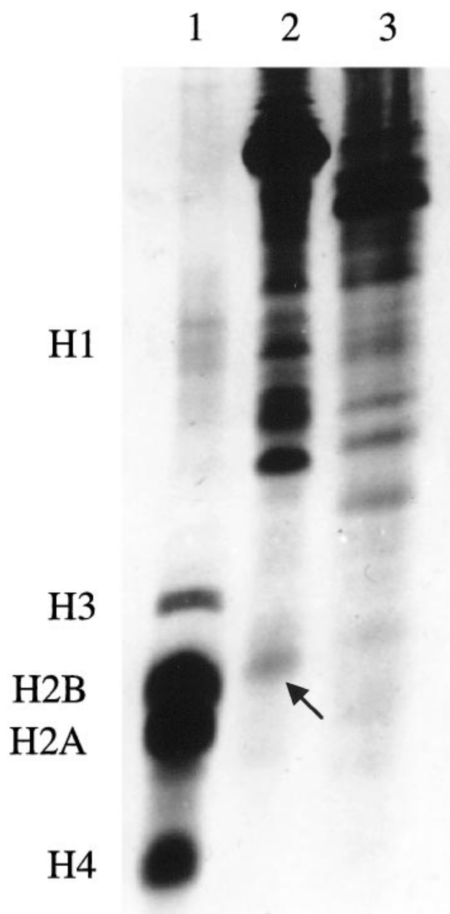

FIG. 1.—Electrophoresis of proteins extracted from *Drosophila melanogaster.* Lane 1 contains calf thymus histones (H1, H2A, H2B, H3, and H4) as a reference. Lane 2 contains proteins from testes of 285 unmated males (>30 days old). The arrow indicates the protein that was isolated for sequencing. Lane 3 contains total protein from 12 one-day-old males.

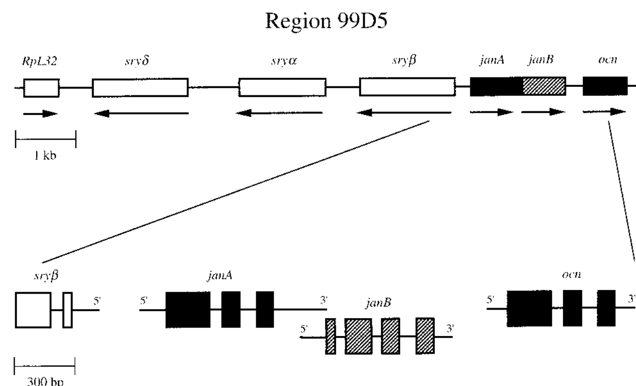

FIG. 2.—Diagram of region 99D5 in *Drosophila melanogaster.* The genomic position of each gene is shown at the top, with arrows indicating the direction of transcription. Below is an enlargement of the *janA, janB,* and *ocn* transcriptional units, with the protein-encoding regions shown as boxes and the introns and untranslated regions of each transcript shown as lines. The 3′ untranslated region (UTR) of *janA* overlaps with the 5′ UTR of *janB* and the start of the *janB* protein-encoding sequence.

Amplification of genomic DNA with primers designed to the ends of the EST sequence (primers oc1 and oc2) resulted in a product approximately 600 bp in length. The DNA sequence of the cloned PCR product was found to be identical to that of the EST with the addition of two small introns (54 and 55 bp) in the genomic sequence. These introns appeared to be homologous to the two introns of *janA* and to the second and third introns of *janB* (fig. 2) because they occurred at identical codon positions within the aligned protein-encoding sequences. The first intron occurred between a first and second codon position; the second occurred between and third and a first codon position.

The cloned PCR product was used as a probe for in situ hybridization to *D. melanogaster* polytene chromosomes, and the *ocn* gene was localized to chromosome 3, region 99D5. Because *janA* and *janB* also map to 99D5, it seemed likely that *ocn* was very close to the *janus* locus and may have resulted from a tandem duplication of *janB.* Using a forward PCR primer specific to *janB* (jb1) and a reverse primer specific to *ocn* (oc2), we were able to amplify a segment of genomic DNA, and the resulting product indicated that *ocn* lay only 200 bp distal to *janB* (fig. 2). This was confirmed by direct sequencing of the *janB-ocn* intergenic region. Our sequence of the *janB* 3′ flanking region agrees with that previously reported by Yanicostas, Vincent, and Lepesant (1989) except for the final 30 bases of their sequence (GenBank accession number M27033), which differ substantially. The recently completed *D. melanogaster* genome sequence (Adams et al. 2000) includes the *janB-ocn* intergenic region and is in agreement with our sequence.

ocn Expression Pattern

Since the *ocn* protein was originally isolated due to its abundance in testis nuclear extracts and shares significant homology with the testis-specific *janB* protein, it seemed likely that *ocn* expression was also specific to testes. To test this hypothesis, we examined the pattern of *ocn* expression using RT-PCR. Our results indicate that *ocn* is expressed in males, but not in females (fig. 3). Males were further dissected into body segments, and a strong band of *ocn* product was obtained from cDNA prepared from the male abdomen (fig. 3). Further dissection of the abdomen into testes and midguts indicated that the abdominal *ocn* expression was specific to the testes (fig. 3). There appeared to be a very faint band corresponding to the expected *ocn* product in the ‘‘head’’ lane of figure 3. This suggests that there may be low levels of *ocn* expression in the male head. The fact that ESTs corresponding to the *ocn* sequence were identified from a combined male and female head cDNA library by the BDGP (Rubin et al. 2000) supports this possibility. A band of a different size appears in the ‘‘thorax’’ lane of figure 3. We believe this represents a nonspecific product, as the band is not distinct and was not present in other amplifications from the male thorax (not shown).

There is a high level of conservation between the 5′ untranslated regions (UTRs) of *janB* and *ocn* (fig.

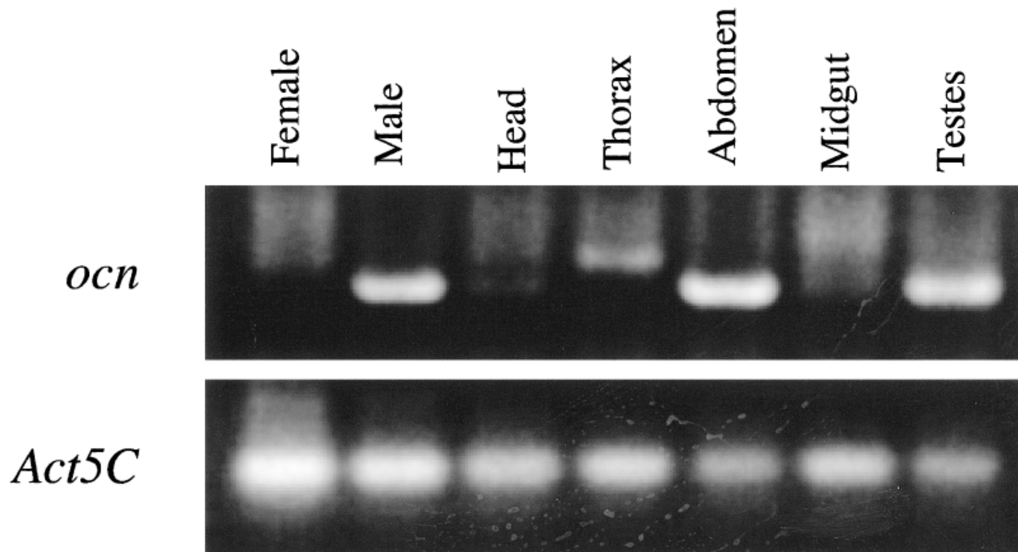

FIG. 3.—Testis specificity of *ocn* transcription demonstrated by RT-PCR. cDNA was synthesized from total RNA purified from adult females and males or from body parts of dissected males. *ocn* cDNA was successfully amplified from whole males, male abdomens, and male testes, but not from females, male heads, male thoraxes, or male midguts. *Act5C* cDNA was amplified from each sample as a control.

4). The *janB* 5′ UTR has previously been shown to contain *cis*-regulatory elements that restrict *janB* translation to the postmeiotic stages of sperm development (Yanicostas and Lepesant 1990). The *janB* 5′ UTR sequence also shows sequence similarity to the 5′ UTR of the sperm-specific gene, *mst(3)gl-9* (Yanicostas, Vincent, and Lepesant 1989), and to the translational control element (TCE) consensus sequence (ACATNAAATTT) common to the *Mst(3)CGP* gene family (reviewed by Schäfer et al. 1995). In *janB,* a near-perfect match to the TCE can be found if a single gap is allowed (ACACAAATTT) or a match of 7 out of 11 bases can be found without gaps (GCACTAAACCT). The closest match to the TCE consensus in *ocn* is 9 out of 11 bases (TCCTAAAATTT). The strong conservation of the *janB* and *ocn* 5′ UTRs indicates functional constraint and suggests that *ocn* translation is also limited to the postmeiotic stages of spermatogenesis. This is consistent with the localization of *ocn* protein by antibody staining (unpublished data).

Phylogenetic Relationship of *janA, janB,* and *ocn*

To further investigate the molecular evolution and duplication patterns of the *janA, janB,* and *ocn* genes, we cloned and sequenced the *janA-ocn* region from other species of the *D. melanogaster* species subgroup (*D. simulans, D. sechellia, D. mauritiana, D. teissieri, D. yakuba, D. erecta,* and *D. orena*) using our *D. melanogaster* PCR primers. The *janA* gene of *D. sechellia* did not amplify with our primers, and its sequence was not determined. In addition, the *D. pseudoobscura janA* and *janB* sequences (Yanicostas et al. 1995) were obtained from GenBank (accession number S77099), as well as the sequence encoding the *Caenorhabditis elegans* P90861 protein (accession number Z81077), which shows similarity to Drosoph-

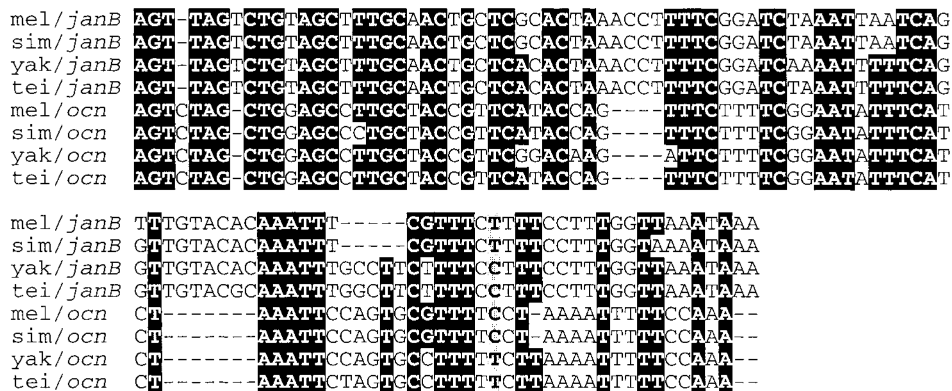

FIG. 4.—Alignment of the *janB* and *ocn* 5′ untranslated regions (UTRs) from *Drosophila melanogaster* (mel), *Drosophila simulans* (sim), *Drosophila yakuba* (yak), and *Drosophila teissieri* (tei). Matches between the two genes are shown in inverse. When two different nucleotides are present at the same position in both genes, they are shown in bold and shaded gray. The *janB* 5′ UTR has been shown to contain translational control elements that restrict translation to the postmeiotic stages of sperm development (Yanicostas and Lepesant 1990).

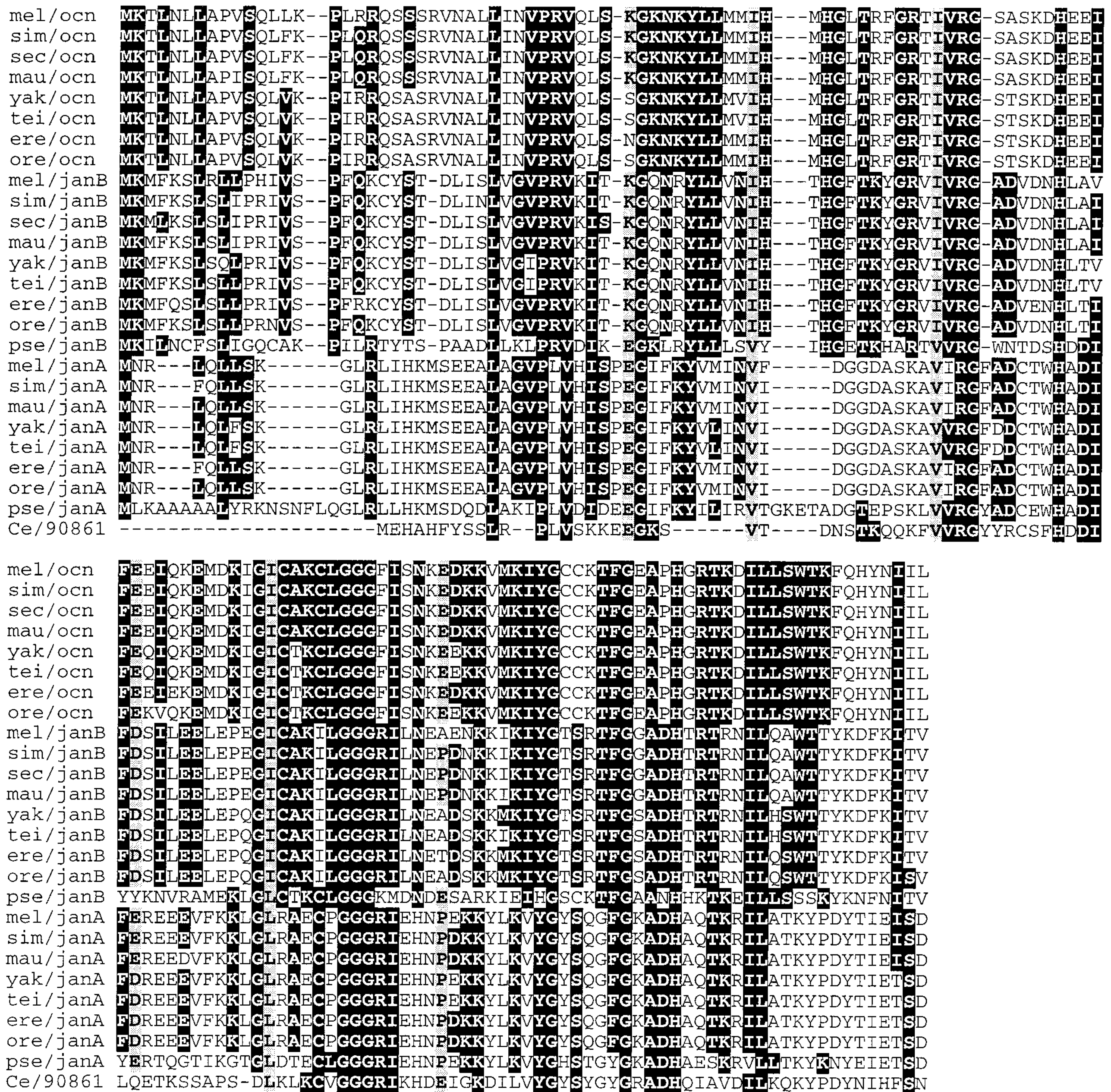

FIG. 5.—Alignment of the predicted *ocn, janB,* and *janA* peptides from *Drosophila melanogaster* (mel), *Drosophila simulans* (sim), *Drosophila sechellia* (sec), *Drosophila mauritiana* (mau), *Drosophila yakuba* (yak), *Drosophila teissieri* (tei), *Drosophila erecta* (ere), *Drosophila orena* (ore), and *Drosophila pseudoobscura* (pse). The *Caenorhibditis elegans* P90861 sequence is shown at the bottom. Residues conserved between two or more genes are shown in inverse. When two different residues are present at the same position in multiple genes, they are shown in bold and shaded gray. Several residues at the C-terminal end (nine from *ocn,* two from *janB,* and three from *janA*) are not shown.

ila *janA*. The aligned protein-encoding sequences (fig. 5) were used to construct a gene tree (fig. 6). The alignment revealed several motifs that are highly conserved. For example, the motif XXRG (where X represents either valine or isoleucine) at residues 67–70 in the alignment was present in all three genes in Drosophila and was also found in the *C. elegans* P90861 protein. Similarly, the tri-glycine motif (GGG) at residues 119–121 was perfectly conserved among all of the sequences in the alignment. This suggests strong selective constraints against changes at these residues and an important functional role for these motifs. Their specific function, however, remains unknown.

Rates of Evolution Within the *D. melanogaster* Species Subgroup

It has been suggested that genes with reproductive function evolve more rapidly (as measured by the ratio ω of the nonsynonymous substitution rate [d_N] to the synonymous substitution rate [d_S], where $\omega = d_N/d_S$) than genes that have no reproductive role (Civetta and

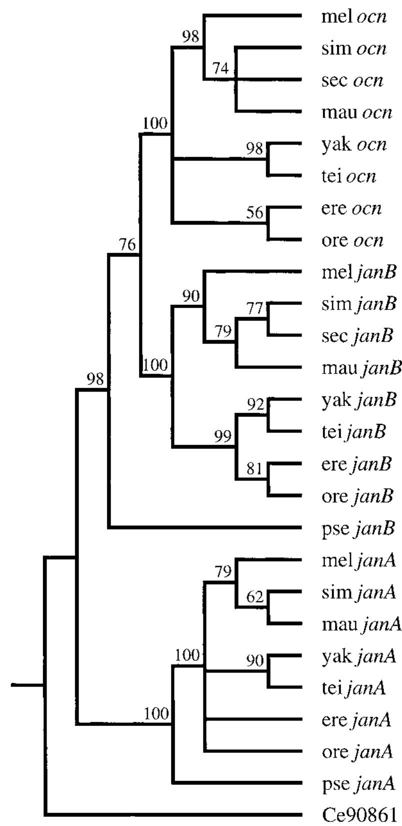

FIG. 6.—Gene tree of *ocn, janA,* and *janB* based on protein-encoding sequences. Species abbreviations are the same as in figure 5. Shown is the 50% majority-rule consensus parsimony tree determined using PAUP* (Swofford 2000). The tree was rooted with the *Caenorhibditis elegans* P90861 sequence. Bootstrap values are shown at each node. Tree optimization by either distance or maximum likelihood produced identical topologies.

Singh 1998). In some cases, such as that of the Drosophila male accessory gland protein gene *Acp26Aa,* a value of $\omega > 1$ has been used to infer the past action of positive selection (Tsaur and Wu 1997). The ω values for *janA, janB,* and *ocn* are all much less than 1 (table 1), so we find no evidence for positive selection by this criterion. The requirement of $\omega > 1$, however, is a strict test for positive selection that assumes strong diversifying selection over many amino acid sites and over many lineages. For this reason, we analyzed our sequence data under various models that allowed for heterogeneous substitution rates among amino acid sites, among lineages, or among genes.

First, we tested four models of selective constraint on amino acid sites within proteins. These models are described in detail in Yang et. al. (2000); we provide a brief summary in table 2. The results of our analyses are shown in table 3. In general, the models can be compared by their likelihood values; a greater likelihood indicates a better fit to the data. In the case of nested models, significance may be tested by comparing twice the likelihood difference ($2\Delta\lambda$) with the critical value of the χ^2 distribution, with the degrees of freedom (df) equal to the difference in the number of parameters between the two models. In all cases, M2 and M3 provided the best fit to the data (table 3). However, these models offered a significant improvement over M0 only in the case of *ocn* (M2 vs. M0: $2\Delta\lambda = 13.4$, df = 2, $P < 0.002$; M3 vs. M0: $2\Delta\lambda = 13.4$, df = 4, $P < 0.01$). For *ocn,* M2 did not offer a significantly better fit than the simpler M1 ($2\Delta\lambda = 0.9$, df = 2, $P > 0.5$). These results suggest that there is significant heterogeneity in ω over different amino acid sites in *ocn,* with a large fraction of sites having $\omega \approx 0$ and the remaining fraction with $\omega \approx 1$. Yang et al. (2000) also describe several models that assume a more complex distribution of ω among sites. For example, they propose the comparison of M7 (beta) and M8 (beta and ω) as a test for positive selection. Likelihood ratio tests based on these models did not provide support for positive selection for any of the three genes (*janA*: $2\Delta\lambda = 0.3$, df = 2, $P > 0.5$; *janB*: $2\Delta\lambda = 0.2$, df = 2, $P > 0.5$; *ocn*: $2\Delta\lambda = 0.1$, df = 2, $P > 0.5$).

The second test we implemented allowed for heterogeneous ω among lineages within a gene. The null model, M0, was identical to the null model for ω variation among sites because it also assumed a constant ω among all lineages. The alternative model, designated "free-ratio" (Yang 1998), allowed for a separate ω along each lineage in the tree topology. Under this model, the number of parameters is equal to the number of branches in the tree. For all three genes, the free-ratio model did not provide a significantly better fit to the data than M0 (*janA*: $2\Delta\lambda = 10.8$, df = 11, $P > 0.25$; *janB*: $2\Delta\lambda = 13.8$, df = 13, $P > 0.25$; *ocn*: $2\Delta\lambda = 14.7$, df = 13, $P > 0.25$). Thus, we found no evidence for variation in ω among lineages. For some of the *janB* and *ocn* branches, the estimated ω was quite large ($\omega \gg 1$; table 3). This occurred for branches that have very few substitutions and was probably highly inaccurate due to the small sample size. For example, in *janB,* ω

Table 1
Summary Statistics

Gene	s	n	κ	S	d_N	d_S	ω	ENC	CAI
janA. . .	7	135	1.53	0.467	0.0161	0.2258	0.0712	48.70	0.0380
janB. . .	8	140	1.98	0.792	0.0338	0.3026	0.1117	55.35	0.0229
ocn. . . .	8	139	2.37	0.570	0.0214	0.2288	0.0935	54.82	0.0251

NOTE.—Values are for species of the *Drosophila melanogaster* species subgroup only. s = number of sequences; n = number of codons; κ = transition/transversion rate; S = total tree length; d_N = nonsynonymous substitution rate; d_S = synonymous substitution rate; $\omega = d_N/d_S$; ENC = effective number of codons (Sharp and Li 1987); CAI = codon adaptation index (Wright 1990). d_N and d_S values represent the averages for all pairwise species comparisons. ENC is averaged over all species. CAI is for *D. melanogaster* only.

Table 2
Summary of Models

Model	Site	Line-age	Gene	Description
Individual				
M0	C	C	V	Null model; performed separately for each gene
M1	V	C	V	''Neutral''; one fraction of sites, p_0, with $\omega = 0$; the remaining fraction, p_1, with $\omega = 1$
M2	V	C	V	''Selection''; same as M1, but with an additional fraction, p_2, of sites with ω_2
M3	V	C	V	''Discrete''; three classes of sites, p_0, p_1, and p_2, with ω_0, ω_1, and ω_2
Free-ratio	C	V	V	Separate ω estimated for each lineage
Combined				
M0	C	C	C	Null model; performed on combined data from all three genes
Gene.	C	C	V	Same as combined M0, except ω estimated separately for each gene

NOTE.—Models M0, M1, M2, and M3 correspond to those of Yang et al. (2000). ''Free-ratio'' corresponds to the model of Yang (1998). The ratio of the nonsynonymous to synonymous substitution rates (ω) is assumed to be either constant (C) or variable (V) among amino acid sites within each protein, among lineages of each gene, or among genes.

along the branch leading to *D. sechellia* was estimated at 89.0 based on two nonsynonymous changes and zero synonymous changes.

Finally, we tested for the presence of heterogeneous ω among genes. In this case, the null model, M0, was applied to the combined data of all three genes. This model assumed a constant ω for all sites, for all lineages, and for all genes. The alternative model allowed each gene to have a different ω, estimated by applying M0 to each gene separately under the same parameters used for the combined data. Since the individual gene likelihoods were additive, their sum could be compared with the likelihood of the combined data. Because the *D. sechellia janA* sequence was unavailable, we assumed it to be identical to that of *D. simulans.* This was a conservative assumption for the purposes of our test, because it assumed a branch length of zero between *D. sechellia* and *D. simulans* with an equal number of synonymous and nonsynonymous substitutions (zero). Our results indicated that the gene-specific ω model offers a significantly better fit to the data than M0 ($2\Delta\lambda = 57.1$, df = 2, $P < 0.001$). Thus, we conclude that there is highly significant variation in ω among the *janA, janB,* and *ocn* genes. The model of a separate ω for each gene was also significantly better than models in which any two of the three genes shared a common ω (data not shown).

To compare the rates of evolution of the *janA, janB,* and *ocn* genes with those of genes that do not have reproductive function, we calculated ω for other protein-encoding genes with sequences available from species throughout the *D. melanogaster* species subgroup: Cu-Zn superoxide dismutase (*Sod*), alpha-amylase proximal (*Amy-p*) and distal (*Amy-d*), and *Adh.* All four of these genes encode proteins with metabolic function, and all four have smaller ω values than *janA, janB,* or *ocn* (fig. 7). In all cases, ω was significantly lower than that of *janA* as determined by the test described above (*Sod*:

Table 3
Likelihood Values and Parameters Under Models of Heterogeneous ω Within and Among Genes

Gene	Model[a]	λ	Estimated Parameters
janA	M0	−837.47	$\omega = 0.071$
	M1	−837.69	$p_0 = 0.883$, $p_1 = 0.117$
	M2	−835.76	$p_0 = 0.776$, $p_1 = 0.000$, $p_2 = 0.224$, $\omega_2 = 0.330$
	M3	−835.77	$p_0 = 0.297$, $p_1 = 0.483$, $p_2 = 0.220$, $\omega_0 = 0.001$, $\omega_1 = 0.001$, $\omega_2 = 0.334$
	Free-ratio	−832.06	ω estimated for each of 12 branches (range = 0.296–0.001)
janB	M0	−1,069.78	$\omega = 0.112$
	M1	−1,078.07	$p_0 = 0.762$, $p_1 = 0.238$
	M2	−1,069.14	$p_0 = 0.426$, $p_1 = 0.000$, $p_2 = 0.547$, $\omega_2 = 0.198$
	M3	−1,069.14	$p_0 = 0.430$, $p_1 = 0.047$, $p_2 = 0.523$, $\omega_0 = 0.001$, $\omega_1 = 0.199$, $\omega_2 = 0.199$
	Free-ratio	−1,062.86	ω estimated for each of 14 branches (range = 89.0–0.001
ocn	M0	−921.48	$\omega = 0.094$
	M1*	−915.25	$p_0 = 0.866$, $p_1 = 0.134$
	M2*	−914.78	$p_0 = 0.000$, $p_1 = 0.094$, $p_2 = 0.906$, $\omega_2 = 0.021$
	M3*	−914.76	$p_0 = 0.111$, $p_1 = 0.644$, $p_2 = 0.245$, $\omega_0 = 0.861$, $\omega_1 = 0.015$, $\omega_2 = 0.015$
	Free-ratio	−914.14	ω estimated for each of 14 branches (range = 89.0–0.001)
Combined[b].	M0	−2,857.95	$\omega = 0.096$
	Gene**	−2,829.40	$\omega_1 = 0.075$, $\omega_2 = 0.111$, $\omega_3 = 0.088$

[a] Models are the same as in table 2. All tests of significance were performed against M0.

[b] In the combined analysis, the transition/transversion rate, κ, was fixed at 1.90 (as estimated from the combined data). For this reason, ω values differ slightly from those of the individual genes, and the combined likelihood, λ, differs slightly from the sum of the individual likelihoods.

* $P < 0.01$.

** $P < 0.001$.

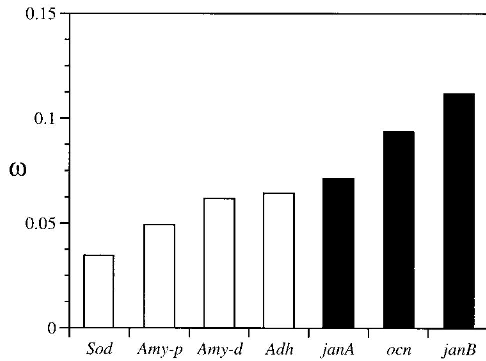

FIG. 7.—Ratio of nonsynonymous to synonymous substitution rates (ω) for protein-encoding genes in the *Drosophila melanogaster* species subgroup. For each gene, ω is averaged over all species. Genes are arranged in order of increasing ω. Open boxes represent genes with metabolic function. Solid boxes represent genes with reproductive function. Average pairwise values of d_N/d_S for *Sod, Amy-p, Amy-d,* and *Adh* are 0.0061/0.1735, 0.0122/0.2489, 0.0142/0.2333, and 0.0134/0.2089, respectively.

$2\Delta\lambda = 42.4$, df = 2, $P < 0.001$; *Amy-p*: $2\Delta\lambda = 108.7$, df = 2, $P < 0.001$; *Amy-d*: $2\Delta\lambda = 87.6$, df = 2, $P < 0.001$; *Adh*: $2\Delta\lambda = 58.2$, df = 2, $P < 0.001$).

Discussion

The sequence similarity among *janA, janB,* and *ocn,* together with their physical proximity, suggests that this gene cluster resulted from two duplication events. We propose a duplication pattern of *janA* → *janA janB* → *janA janB ocn.* Since only one protein with similarity to any of the three Drosophila proteins was identified in the *C. elegans* genome (*C. elegans* Sequencing Consortium 1998), and this sequence shows the greatest similarity to *janA,* it is likely that *janA* most closely resembles the ancestral sequence. The observation that *janA* shows a spatially and developmentally general pattern of expression in both males and females (Yanicostas, Vincent, and Lepesant 1989), while *janB* and *ocn* are expressed only in the male germ line, suggests that *janA* function is also ancestral and that increased specificity of function followed at least the first duplication event. The initial duplication clearly predates the divergence of the *D. melanogaster* and *D. obscura* species groups. The timing of the second duplication event can be inferred from the phylogenetic tree (fig. 6), where the *D. pseudoobscura janB* gene falls outside of the clade containing the *D. melanogaster* species subgroup *janB* and *ocn* genes. This is most parsimonious with a second duplication event that took place after the split of the *D. melanogaster* and *D. obscura* species groups. Because *D. pseudoobscura janB* shares its intron-exon structure with the *D. melanogaster* species subgroup *janB* genes, *ocn* appears to be the most derived. The above model predicts that *ocn* is absent in *D. pseudoobscura.* Consistent with this prediction, we have not been able to isolate an *ocn* homolog from *D. pseudoobscura* by PCR-based methods, nor have we been able to detect homology to *ocn* in the previously published sequence of the *D. pseudoobscura janA-janB* region (Yanicostas et al. 1995). We cannot, however, eliminate the possibility of an *ocn* homolog in *D. pseudoobscura* because at present, only about 200 bp downstream of *janB* have been sequenced. It remains possible that an *ocn* homolog lies further distal to *janB* in *D. pseudoobscura* than in species of the *D. melanogaster* species subgroup.

While the above duplication narrative is most parsimonious with respect to the sequence-based phylogenetic tree, it requires two separate instances of intron gain/loss, because all of the *janB* sequences (including that of *D. pseudoobscura*) contain a first intron that is not present in any of the *janA* or *ocn* sequences. One possibility is that an intron inserted into the ancestral *janB* gene following the first duplication event and was lost in the *ocn* lineage following the second duplication. Alternatively, the first intron might have been present in the ancestral *janA* sequence and subsequently lost in both *janA* and *ocn* following duplication. In Drosophila, several examples indicate that parallel loss of introns may be relatively common (Anderson, Carew, and Powell 1993; Da Lage, Wegnez, and Cariou 1996). Our tree is also consistent with two intron insertions in the *D. pseudoobscura* and *D. melanogaster* species subgroup *janB* gene lineages, but we consider this less likely because it requires parallel intron insertion into the same location of the coding region. Since the 3′ end of the *janA* transcript overlaps with the 5′ end of *janB* (fig. 2), it is likely that the *janB* sequence faces additional selective constraints required for proper processing of the *janA* transcript. Such constraints would not apply to *ocn* and may account for the divergence in intron-exon structure between *janB* and *ocn.*

Although none of the proteins encoded by *janA, janB,* or *ocn* contain recognizable structural motifs that suggest a molecular function, it is likely that they are involved in chromatin packaging in sperm nuclei. The predicted molecular weights of the *janA, janB,* and *ocn* proteins in *D. melanogaster* are 15.22, 15.86, and 16.89 kDa, respectively. All three proteins are basic, containing 18%–21% positively charged amino acid residues. The fraction of basic residues in each protein is similar to that found in Drosophila histones. In addition, the *ocn* protein shows a migration pattern similar to histones H2 and H3 when separated by gel electrophoresis (fig. 1). Both *ocn* mRNA and protein are abundant in the testes of mature males, and the strong conservation between the *janB* and *ocn* 5′ UTRs suggests that *ocn* translation is restricted to the postmeiotic stages of spermatogenesis.

Maximum-likelihood testing of different models of selective constraint indicates that rates of synonymous and nonsynonymous substitution within the *janA-ocn* region are best explained by assigning a different ω parameter to each of the three genes. There is strong statistical support for differences in ω among all three genes, with the highest rate in *janB* and the lowest in *janA.* Since ω is the ratio of d_N to d_S, differences in ω among genes may result from differences in either of these quantities. For example, the higher ω of *janB* could be the result of either an increased d_N or a decreased d_S relative to *janA* and *ocn.* Our results indicate that the differences in ω among *janA, janB,* and *ocn* are due primarily to differences in d_N (table 1). Furthermore,

if the increased ω was the result of reduced synonymous substitution rates due to purifying selection on synonymous codon sites, we would expect genes with higher ω to show greater levels of codon bias. Two measures of codon bias reveal that there is no positive correlation between ω and codon bias (table 1). In fact, we find the opposite: genes with higher values of ω show lower levels of codon bias. Thus, we conclude that the differences in ω among these genes are not the result of reduced rates of synonymous substitution caused by increased purifying selection against unpreferred codons.

The likelihood analyses presented in table 3 assume a tree topology of the *D. melanogaster* species subgroup identical to that of the *janB* gene in figure 6. This tree contains no polytomies, and each node is supported by a bootstrap of ≥77%. An identical tree is predicted for the *D. melanogaster* species subgroup when data from the three genes are combined. Tree topologies based on *ocn* and *janA,* however, are more ambiguous and contain several polytomies (fig. 6). To investigate whether the tree topology might affect our results, we repeated the above analyses using topologies suggested by either *ocn* or *janA*. Consistent with previous reports (Yang et al. 2000), we found that the use of reasonable, alternate tree topologies had a negligible effect on the results and did not alter their statistical significance (not shown).

Comparison of genes from a group of closely related species, such as the *D. melanogaster* species subgroup used here, allows for powerful statistical tests to detect differences in the rates of evolution among genes. Using this method, we find significant differences in evolutionary rates both among paralogous genes located within a 2.4-kb region of the genome and between genes with reproductive and metabolic function. Although the sample size is small due to the lack of sequence availability for many species of the *D. melanogaster* species subgroup, these results are consistent with those of Civetta and Singh (1998), who found a higher ratio of nonsynonymous to synonymous substitution rates in genes with a sex-related function than in genes with developmental or metabolic function. Our results lend support to the ''faster male'' hypothesis (Wu and Davis 1993) and suggest that rapid evolution of genes involved in male fertility may play a role in reproductive isolation between species. The increased rate of molecular evolution observed for *janA, janB,* and *ocn* could be the result of either positive selection for amino acid replacements or relaxed selective constraints. It is generally not possible to distinguish between these two possibilities through interspecific comparison of protein-encoding sequences. One exception to this occurs when there is strong diversifying selection, such as with antigenic proteins of infectious pathogens (Fitch et al. 1997; Yang et al. 2000), resulting in a value of ω significantly greater than 1. Such cases, however, appear to be quite rare (Endo, Ikeo, and Gojobori 1996). Further analysis of patterns of intraspecific variation in the *janA-ocn* region will allow the application of more powerful statistical tests (e.g., Hudson, Kreitman, and Aguadé 1987; McDonald and Kreitman 1991) to determine if positive selection has played a role in the molecular evolution and functional divergence of this duplicated gene family.

Supplementary Material

The sequence of the *D. melanogaster ocn* gene has been submitted to the GenBank database under accession number AF231190. Sequences of *janA, janB,* and *ocn* from other species of the *D. melanogaster* species subgroup have been submitted under accession numbers AY013339–AY013344, AY013345–AY013351, and AY013352–AY013358, respectively.

Acknowledgments

We thank members of the Hartl lab for their suggestions and advice. Two anonymous reviewers provided valuable comments on a previous version of the manuscript. This research was supported by National Institutes of Health grants GM60035 and HG01250 to D.L.H.

LITERATURE CITED

Adams, M. D., S. E. Celniker, R. A. Holt et al. (92 coauthors). 2000. The genome sequence of *Drosophila melanogaster.* Science **287**:2185–2195.

Albalat, R., G. Marfany, and R. Gonzalez-Duarte. 1994. Analysis of nucleotide substitutions and amino acid conservation in the *Drosophila Adh* genomic region. Genetica **94**:27–36.

Anderson, C. L., E. A. Carew, and J. R. Powell. 1993. Evolution of the *Adh* locus in the *Drosophila willistoni* group: the loss of an intron, and shift in codon usage. Mol. Biol. Evol. **10**:605–618.

Ashburner, M. 1998. Speculations on the subject of alcohol dehydrogenase and its properties in *Drosophila* and other flies. BioEssays **20**:949–954.

C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science **282**:2012–2018.

Civetta, A., and R. S. Singh. 1998. Sex-related genes, directional sexual selection, and speciation. Mol. Biol. Evol. **15**: 901–909.

Da Lage, J. L., M. Wegnez, and M. L. Cariou. 1996. Distribution and evolution of introns in Drosophila amylase genes. J. Mol. Evol. **43**:334–347.

Endo, T., K. Ikeo, and T. Gojobori. 1996. Large-scale search for genes on which positive selection may operate. Mol. Biol. Evol. **13**:685–690.

Fitch, W. M., R. M. Bush, C. A. Bender, and N. J. Cox. 1997. Long term trends in the evolution of H(3) HA1 human influenza type A. Proc. Natl. Acad. Sci. USA **94**:7712–7718.

Force, A., M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151**: 1531–1545.

Gusse, M., P. Sautière, D. Bélaiche, A. Martinage, C. Roux, J.-P. Dadoune, and P. Chevaillier. 1986. Purification and characterization of nuclear basic proteins of human sperm. Biochim. Biophys. Acta **884**:124–134.

Hudson, R. R., M. Kreitman, and M. Aguadé. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics **116**:153–159.

Krakauer, D. C., and M. A. Nowak. 1999. Evolutionary preservation of redundant duplicated genes. Semin. Cell Dev. Biol. **10**:555–559.

Lange, B. W., C. H. Langley, and W. Stephan. 1990. Molecular evolution of the Drosophila metallothionein genes. Genetics **126**:921–932.

McDonald, J. H., and M. Kreitman. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila.* Nature **351**:652–654.

Nurminsky, D. I., M. V. Nurminskaya, D. De Aguiar, and D. L. Hartl. 1998. Selective sweep of a newly evolved sperm-specific gene in *Drosophila.* Nature **396**:572–575.

Ohno, S. 1970. Evolution by gene duplication. Springer-Verlag, Berlin.

Petrov, D. A., E. R. Lozovskaya, and D. L. Hartl. 1996. High intrinsic rate of DNA loss in *Drosophila.* Nature **384**: 346–349.

Presgraves, D. C., and H. A. Orr. 1998. Haldane's rule in taxa lacking a hemizygous X. Science **282**:952–954.

Rubin, G. M., L. Hong, P. Brokstein, M. Evans-Holm, E. Frise, M. Stapleton, and D. A. Harvey. 2000. A *Drosophila* complementary cDNA resource. Science **287**:2222–2224.

Sambrook, J., E. F. Fritsch, and T. Maniatis. 1989. Molecular cloning: a laboratory manual. 2nd edition. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

Schaeffer, S. W., and C. F. Aquadro. 1987. Nucleotide sequence of the *Adh* gene region of *Drosophila pseudoobscura*: evolutionary change and evidence for an ancient gene duplication. Genetics **117**:61–73.

Schäfer, M., K. Nayernia, W. Engel, and U. Schäfer. 1995. Translational control in spermatogenesis. Dev. Biol. **172**:344–352.

Sharp, P. M., and W.-H. Li. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. **15**: 1281–1295.

Swofford, D. L. 2000. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer, Sunderland, Mass.

Theodore, L., A.-S. Ho, and G. Maroni. 1991. Recent evolutionary history of the gene *Mtn* in Drosophila. Genet. Res. **58**:203–210.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25**: 4876–4882.

Tsaur, S.-C., and C.-I. Wu. 1997. Positive selection and the molecular evolution of a gene of male reproduction, *Acp26Aa* of *Drosophila.* Mol. Biol. Evol. **14**:544–549.

Wright, F. 1990. The 'effective number of codons' used in a gene. Gene **87**:23–29.

Wu, C.-I., and A. W. Davis. 1993. Evolution of postmating reproductive isolation: the composite nature of Haldane's rule and its genetic bases. Am. Nat. **142**:187–212.

Yang, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. **15**:568–573.

———. 2000. Phylogenetic analysis by maximum likelihood (PAML). Version 3.0. University College London, London.

Yang, Z., R. Nielsen, N. Goldman, and A-M. K. Pederson. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155**:431–449.

Yanicostas, C., P. Ferrer, A. Vincent, and J.-A. Lepesant. 1995. Separate *cis*-regulatory sequences control expression of *serendipity* β and *janus A,* two immediately adjacent *Drosophila* genes. Mol. Gen. Genet. **246**:549–560.

Yanicostas, C., and J.-A. Lepesant. 1990. Transcriptional and translational *cis*-regulatory sequences of the spermatocyte-specific *Drosophila janusB* gene are located in the 3′ exonic region of the overlapping *janusA* gene. Mol. Gen. Genet. **224**:450–458.

Yanicostas, C., A. Vincent, and J.-A. Lepesant. 1989. Transcriptional and posttranscriptional regulation contributes to the sex-regulated expression of two sequence-related genes at the janus locus of *Drosophila melanogaster.* Mol. Cell. Biol. **9**:2526–2535.

David M. Rand, reviewing editor

Accepted January 19, 2001